Institute for Youth in Policy

Policy Media Department | Brief

5214F Diamond Heights Blvd #3055
San Francisco, CA 94109
+1 (509) 565-0545
eic@yipinstitute.org
www.yipinstitute.org

Regulating AI Detection in Academic Work

Policy Leads: Ava Szajnuk[#1], Bhavana Rupakula[#2]

Policy Analysts: Malek Kamal[#3], Queen-Asset Blissett[#4], Lola Agboola[#5], Siddi J.[#6], Gautham Parvathanenil[#7], Lachlan Feichthaler[#8]

Technology Policy, Institute for Youth in Policy

[1]avaszajnuk@yipinstitute.org, [2]bhavana@yipinstitute.org

***Abstract* — The increasing use of generative artificial intelligence in education presents new challenges for academic institutions attempting to distinguish between legitimate AI-assisted learning and academic dishonesty. The absence of clear institutional policies, untrustworthy AI detection mechanisms, and unclear requirements for AI disclosures leave room for ambiguity. The proposed policy brief outlines steps that educational institutions can undertake in order to create clear and coherent rules for the use of AI in academic work, mandate disclosures when AI makes substantial contributions to student work, avoid relying only on automatic AI detection software, and have human-based review and appeal processes in case of any violations.**

***Keywords* — Artificial Intelligence (AI); generative AI; academic integrity; higher education; AI detection; academic misconduct; transparency; due process; education policy.**

I. Executive Summary

Schools and universities are turning to artificial intelligence detection tools in greater numbers to identify AI-produced work, yet the speed with which they have been embraced has left a void in the standards that ought to regulate them. While the aim is to uphold academic integrity, the truth is these systems offer only probabilistic readings, not hard and fast answers. There is a risk of original writing being wrongly flagged, especially from multilingual students or those whose style is more formulaic. All too often an institution will take a detection score as evidence of wrongdoing and penalise a student, all without any transparency or chance for the student to put forward a case.

Our findings in this brief are clear: AI detection has its place as an aid to an educator's judgment and a wider assessment of a student's output, but it can never be a substitute. We advise against any disciplinary measure hinging on an automated result alone. Vendors should be made to put their accuracy and bias data on the table, and there must be a standardised appeals procedure so students can see what evidence is being used. In addition, schools would do well to put in place unambiguous policies on AI, give their staff the training to look at drafts and writing history when in doubt, and design assignments that promote the responsible use of such technology. Such measures are necessary to safeguard both the rights of the student and the trust placed in the educational process.

II. Overview

A. Relevance

Since generative tools, like ChatGPT, became widely available in 2022, schools have begun to implement AI detection tools to counteract the use of artificial intelligence by students for cheating. Schools and universities utilize programs like Turnitin and GPTZero to help identify assignments that may have been written by artificial intelligence. While these tools were brought in to support academic integrity,

there are rising concerns about whether the programs' evaluations should be used as evidence for academic misconduct cases. As a result, educators, researchers, and policymakers have begun discussing whether there should be more clarity regarding rules for how AI detection tools are used in schools.

The communities most affected are high school and college students, especially multilingual and international students whose writing styles could be different from the data these AI detection programs are trained on. Recent research has found that some AI detectors are more likely to identify a multilingual speaker's writing as AI-generated, pushing concerns about fairness and bias. As a result, many experts recommend that AI-detection tools should not be the single piece of evidence for disciplinary action. As academic institutions continue to adopt AI technology and the technology becomes more widespread, policymakers will need to balance academic integrity with fairness and no bias.

This issue has affected many students, causing them to face undeserved consequences that interfere with their education. One example is Thierry Rignol, who was suspended after GPTZero falsely accused him of using AI, despite being a high-achieving student on track to become valedictorian. Because Rignol is French, he believed that his formal English writing style and grammar may have caused the AI detector to mistakenly classify his work as AI-generated. While it is impossible to know the exact reason for the false positive, no student should be deprived of their education because of the unreliability of an AI detection tool.

B. Background

Since the rise of AI language models such as ChatGPT, which are capable of producing human-like writing, many people have begun relying on these tools for convenience. One of the largest groups of users is students. According to the College Board's research brief, *College Faculty Perceptions of Generative Artificial Intelligence in Higher Education*, nearly three-quarters (74%) of professors believe that students are using AI to write essays or papers, while 67% believe students use AI to paraphrase or revise their work. Because of these perceptions, many teachers have adopted AI detection websites to determine whether students are using AI, despite the fact that these tools are often inaccurate. At many schools, unauthorized AI use is considered cheating or even plagiarism—offenses that can result in serious disciplinary consequences. However, these accusations are based on unreliable AI detection software that flags students for unfair or misleading reasons.

Many students have become victims of false AI accusations and have received academic penalties despite completing their work honestly. In 2026, multiple cases of false AI accusations have been reported, highlighting the limitations of current AI detection technology. Studies have also shown that AI detectors are more likely to incorrectly flag writing by international English writers, neurodivergent students, and students in higher education. As AI becomes more common in education, schools should develop more reliable methods of identifying AI-generated work rather than relying on inaccurate detection websites. Academic integrity should be protected through evidence-based investigations and fair reasoning instead of software that can make costly mistakes.

C. Notable Stakeholders

Many groups have a stake in these policies. Students want to ensure that their work is graded fairly and that they are not wrongly accused of cheating because of a program's inefficiency and an inaccurate AI detection score. Teachers and schools want to promote an academically honest environment while still treating the students fairly. AI detection companies also have an interest because new policies would require them

to be more transparent with how their tools work and how accurate they are. Parents and state education agencies are also involved because they want to uphold the schools' academic standard while still protecting students' rights.

III. Policy Problem

With the rising prevalence of artificial intelligence in academic work, educators and administrators alike have taken to using artificial intelligence itself as a tool to fight against schoolwork using AI. However, this method has had mixed results for students, particularly those in middle school, high school, and college, who regularly submit essays, research papers, and written assignments through plagiarism or AI-detection software. AI systems are not perfectly accurate, and can incorrectly flag student work as AI-generated. This places students in a difficult position, particularly non-native English speakers, who may potentially face academic penalties despite completing work honestly.

These false accusations can have significant academic and emotional implications. For example, a student applying to college who has their essay flagged may potentially face disciplinary action, losing out on the opportunity to pursue higher education. Likewise, students who use grammar-checking software or accessibility tools to improve clarity may find that their legitimate work receives a higher AI-generated score, despite the fact that these tools are widely accepted in educational settings. Even tools like Grammarly, which are approved at many institutions, come back on AI detector checks, leaving students fearful of the possibility of having their work flagged. As a result, students may feel overwhelming anxiety and stress, leading to mistrust in teachers and school administrative processes when AI detectors are used as definitive evidence. Instead of spending time working on stronger writing and critical thinking skills, students focus on how to ensure they aren't flagged for submitting illegitimate work.

Therefore, policymakers should seek to regulate the use of AI detection systems, by emphasizing transparency, forcing accurate figures out of companies, and requiring models to be trained on a wide and diverse dataset, to ensure marginalized groups are not unfairly discriminated against. Meanwhile, schools should be offering students the chance to appeal, a vital part of protecting students' rights, while simultaneously upholding academic integrity. After all, even 1% of essays flagged results in a significant amount of false accusations, endangering the future of the youth.

IV. Policy Options

Online platforms with algorithmic recommendations and a reasonable probability of being accessed by minors must implement appropriate mechanisms to detect, disclose, and mitigate the spread of AI-generated content. In addition to using commercially reasonable methods to detect AI-generated images, videos, audio, and text from uploaded content, any undetected AI content is to be labeled with an appropriate notice before it is viewed, shared, or recommended. The recommendation algorithms must also be designed to mitigate the spread of AI-generated content, especially when the content may be accessed by minors.

In order to prevent automated systems from taking independent final decisions on reports, covered platforms are required to create a Human-in-the-Loop Review System for reports regarding AI content, including deepfakes, manipulation, and algorithmic amplification of such content which could cause harm to minors. Such reports will be reviewed by human moderators pre-trained in issues related to digital safety and protection of children from misinformation. These moderators may issue

requests for corrective disclosures, limit the promotion of AI content algorithms, remove the content that violates this Act, or restore wrongly removed content. Platforms must provide an easily accessible mechanism for appeals and produce yearly transparency reports on the number of AI content reports, their outcomes, the response time, and disclosure compliance rate.

V. Impact On Young People

A critical aspect of discussing the regulation of AI detection in academic work is potential impacts on the youth members of our nation. As elaborated on earlier in this brief, students are increasingly turning to AI within academic spaces, and this growing development is fundamentally changing the landscape of school in its entirety.

The first facet of youth impact regarding academic AI covers the potential emotional impacts on our adolescent population. In a 2025 National Library of Medicine study titled "Emotional Profiles & Their Relationship with the Use of Artificial Intelligence in University Students", researchers consider the AI's use as an informational, academic, and emotional support hub. The study also explores where there are "statistically significant differences between the identified emotional intelligence (EI) profiles and the purposes for which AI is used." Ultimately, the NIH examined the association between emotional intelligence and the use of artificial intelligence. The results of the study indicated that while AI aids efficiency and personalized learning in youth, it also poses risks of cognitive dependency and reduced critical thinking, providing recommendations that include treating AI as a "supportive tool" rather than a replacement for cognitive effort in order to protect academic integrity. This shows that academic AI use can cognitively and emotionally affect our student population if a dependency is formed, and if there is no intervention from a legal or educational authority.

Another facet of academic AI and its impact on youth is the power of regulation through policymaking or other rule-enacting organization. The Brookings Institution released a commentary in January of 2026, leading with a powerful title: "AI's Future for Students is in Our Hands." Its authors pose important questions about the duty and role our authorities have in shaping the future of student learning and development, including whether AI will substantially improve children's education or present risks that undermine it. "As educators contemplate AI's integration into classroom practice," Mary Burns and Rebecca Winthrop ask as members of Brookings' global taskforce on AI in education, "how can we embrace its transformational potential while minimizing risks to student agency, deep learning, and emotional well-being?" According to this commentary, the Brookings Institution's Center for Universal Education has been investigating these very questions since September of 2024; they have done so by holding interviews with hundreds of educators and parents as well as students, by consulting with leaders in education and technology, and by examining more than 400 research articles. The results begin by emphasizing the potential benefits of AI in education, extending to teachers by reducing time spent on numerous teaching-related tasks; AI can allow teachers to focus on individualized student attention while "enhancing curriculum and instruction." In a similar vein, AI can "empower student learning by providing access to otherwise unavailable learning opportunities." One aspect that Brookings especially highlights is the way AI can make learning more engaging and accessible for students with disabilities, neurodivergent learning pathways, and multilingual status. However, according to the same article, "AI's risks currently overshadow its

benefits." The threat of AI to students, particularly unsupervised use, are primarily "cognitive, emotional, and social." AI tools prioritize efficiency rather than accuracy in learning and well-being by performing inconsistently across assigned tasks and confidently presenting misinformation as the truth.

The article outlines many pitfalls of AI in education, but the general takeaway speaks to the previously-ordained topic of this section: the power of regulation. Brookings research concludes that AI's "educational evolution" is in the hands of our individuals as well as our institutions; "technology companies, governments, education systems, civil society, teachers, parents, and students themselves all play a role in mitigating AI's risks and harnessing the benefits." Primary suggestions for regulatory authorities include ethical and trustworthy AI design spearheaded by protections embedded during the design phase, responsible governing practices with strong regulatory frameworks, and encouraging adults to model healthy technology use in the home.While 31 states have published guidance or policies for AI in K-12 education as of December 2025, there is yet to be a national framework of rules for academic AI use. However, we do have current examples and suggestions for how to properly execute AI regulations, spearheaded by the rest of the world. The European Union's Artificial Intelligence Act employs a "risk-based approach that bans unacceptable threats, mandates transparency for limited-risk systems, and regulates high-risk applications." The act does so while also requiring protections for user privacy and enacting age limits for "adult-oriented AI." By emphasizing these pillars, the act seeks to impact youth populations by emphasizing emotional privacy, ensuring fair educational opportunities, and changing how adolescents interact with daily digital tools, ultimately preventing youth from taking advantage of AI as well as vice-versa.

VI. Conclusion

AI-detection tools may assist educators in identifying work that warrants further review, but their results should not be treated as definitive evidence of academic dishonesty. Because these systems remain probabilistic and may produce false positives, particularly across diverse writing styles and student populations, educational institutions need clear standards governing their use.

It is important for schools and universities to prohibit disciplinary decisions based solely on automated detection scores, establish transparent policies regarding permitted and prohibited uses of generative AI, require meaningful human review of suspected violations, and provide students with access to evidence and fair appeal procedures. Institutions should also require greater transparency from AI-detection vendors regarding the accuracy, limitations, and potential biases of their systems.

A balanced approach does not require schools to abandon academic integrity enforcement or ignore the challenges created by generative AI. Instead, it requires institutions to recognize the limits of automated detection and ensure that efforts to protect academic integrity do not undermine fairness, transparency, and due process. By treating AI detection as one source of information rather than a final judgment, educational institutions can better protect both student rights and the credibility of academic assessment.

VII. Acknowledgement

The Institute for Youth in Policy wishes to acknowledge Adwaya Yesare for editing this policy brief.

VIII. References

Anderson, Nate. 'How a Yale AI-Cheating Dispute https://arstechnica.com/tech-policy/2026/07/how-a-yale-ai-cheating-dispute-became-a-13-count-federal-lawsuit/.

Deep, Promethi Das, et al. 'Evaluating the Effectiveness and Ethical Implications of AI Detection Tools in Higher Education'. *Information*, vol. 16, no. 10, Oct. 2025, p. 905, https://doi.org/10.3390/info16100905.

Giray, Louie, et al. 'Beyond Policing: AI Writing Detection Tools, Trust, Academic Integrity, and Their Implications for College Writing'. *Internet Reference Services Quarterly*, Jan. 2025, pp. 1–34, https://doi.org/10.1080/10875301.2024.2437174.

'Guidance for Generative AI in Education and Research'. *Unesco.Org*, UNESCO, 7 Sept. 2023, https://www.unesco.org/en/articles/guidance-generative-ai-education-and-research?hub=195885&.

'Guides: Generative AI Detection Tools: The Problems With AI Detectors: False Positives and False Negatives'. *Sandiego.Edu*, 2023, https://lawlibguides.sandiego.edu/c.php?g=1443311&p=10721367.

Hirsch, Amanda. 'AI Detectors: An Ethical Minefield - Center for Innovative Teaching and Learning'. *Center for Innovative Teaching and Learning* -, 12 Dec. 2024, https://citl.news.niu.edu/2024/12/12/ai-detectors-an-ethical-minefield/.

Maphoto, Kgabo Bridget, et al. 'Automated Integrity or Automated Injustice? Turnitin's AI Detection in Multilingual Writing Contexts'. *Assessing Writing*, vol. 69, June 2026, p. 101086, https://doi.org/10.1016/j.asw.2026.101086.

'New College Board Research: Faculty Express Near-Universal Concern That Student AI Use Undermines Original Writing and Critical Thinking'. *Collegeboard.Org*, 2026, https://newsroom.collegeboard.org/new-college-board-research-faculty-express-near-universal-concern-student-ai-use-undermines.

Pindell, Nate. 'The Challenge of AI Checkers | Center for Transformative Teaching | Nebraska'. *Unl.Edu*, 2024, https://teaching.unl.edu/ai-exchange/challenge-ai-checkers/.

Thacker, Nirmal. 'AI Cheating Lawsuits Tracker — Every Case, Who Won (2026)'. *GradPilot*, 19 June 2026, https://gradpilot.com/news/ai-cheating-lawsuits-tracker.

Unseen Studio. "Designer Sketching Wireframes." *Unsplash*, June 11, 2015. https://unsplash.com/photos/person-writing-on-brown-wooden-table-near-white-ceramic-mug-s9CC2SKySJM.

'Using the AI Writing Report'. *Turnitin Guides*, 2024, https://guides.turnitin.com/hc/en-us/articles/22774058814093-Using-the-AI-Writing-Report?Became a 13-Count Federal Lawsuit'. *Ars Technica*, 31 July 2026,